

# SOME IMPROVEMENTS ON LOCALLY REPAIRABLE CODES

JUN ZHANG, XIN WANG, AND GENNIAN GE

**ABSTRACT.** The locally repairable codes (LRCs) were introduced to correct erasures efficiently in distributed storage systems. LRCs are extensively studied recently. In this paper, we first deal with the open case remained in [40] and derive an improved upper bound for the minimum distances of LRCs. We also give an explicit construction for LRCs attaining this bound. Secondly, we consider the constructions of LRCs with any locality and availability which have high code rate and minimum distance as large as possible. We give a graphical model for LRCs. By using the deep results from graph theory, we construct a family of LRCs with any locality  $r$  and availability 2 with code rate  $\frac{r-1}{r+1}$  and optimal minimum distance  $O(\log n)$  where  $n$  is the length of the code.

## 1. INTRODUCTION

In distributed storage systems, redundancy should be introduced to protect data against device failures. The simplest and most widespread technique used for data recovery is replication. However, this strategy entails large storage overhead and is nonadaptive for modern systems supporting the “Big Data” environment. To improve the storage efficiency, erasure codes are employed, such as Windows Azure [16], Facebook’s Hadoop cluster [32], where the original data are divided into  $k$  equal-sized fragments and then encoded into  $n$  fragments ( $n \geq k$ ) stored in  $n$  different nodes. It can tolerate up to  $d - 1$  node failures, where  $d$  is the minimum distance of the erasure code. Particularly, the maximum distance separable (MDS) code is a kind of erasure code that attains the optimal minimum distance with respect to the Singleton bound and thus provides the highest level of fault tolerance for given storage overhead. However the MDS code is inefficient when we consider the disk I/O complexity, repair-bandwidth and so on.

To improve this, Gopalan et al. [13], Oggier and Datta [25], and Papailiopoulos et al. [28] introduced the concept of repair locality for erasure codes. The  $i$ th coordinate of a code has repair locality  $r$  if it can be recovered by accessing at most  $r$  other coordinates. In this paper, an LRC is referred to an  $[n, k]$  linear code with all symbol locality  $r$ . When  $r \ll k$ , it greatly reduces the disk I/O complexity for repair.

Considering the fault tolerance level, the minimum distance is also a key metric for LRCs. Gopalan et al. [13] first derived the following upper bound for codes with information locality:

$$(1.1) \quad d \leq n - k + 1 - \left(\left\lceil \frac{k}{r} \right\rceil - 1\right)$$

---

The research of Gennian Ge was supported by the National Natural Science Foundation of China under Grant No. 61171198 and Grant No. 11431003, the Importation and Development of High-Caliber Talents Project of Beijing Municipal Institutions, and Zhejiang Provincial Natural Science Foundation of China under Grant No. LZ13A010001.

which is a tight bound by the construction of pyramid codes [15]. Although the bound (1.1) certainly holds for all LRCs, it is not tight in many cases. Later, in [9, 27], the bound (1.1) was generalized to vector codes and nonlinear codes. In order to consider multiple erasures in local repair, two different models were put forward independently by Prakash et al. [29] and Wang et al. [42].

For simplicity, the LRC that achieves the upper bound (1.1) with equality is called an optimal (maximum) LRC in this paper. The first optimal LRCs for the case  $(r+1)|n$  were constructed explicitly in [39] and [33] by using Reed-Solomon codes and Gabidulin codes respectively. Both constructions were built over a finite field whose size is exponential in the code length  $n$ . In [37] for the same case  $(r+1)|n$  the authors constructed an optimal code over a finite field of size comparable to  $n$  by using specially designed polynomials. This construction can be extended to the case  $(r+1) \nmid n$  with the minimum distance  $d \geq n - k - \lceil \frac{k}{r} \rceil + 1$  which is at most one less than the upper bound (1.1). In [1, 38, 45], the authors generalized this idea to the cyclic codes and algebraic geometry codes.

Recently, Song et al. [35] carefully studied the tightness of the bound (1.1), and left two open cases. Another recent improvement was due to [30] where Prakash et al. showed a new upper bound on the minimum distance for LRCs. This bound relies on a sequence of recursively defined parameters and is tighter than the bound (1.1). But no general constructions attaining this new bound was presented. A great improvement for this problem is made by Wang and Zhang in [40]. The authors carried out an in-depth study of the two problems: what is the largest possible minimum distance for an  $[n, k]$  LRC? How to construct an  $[n, k]$  LRC with the largest possible minimum distance? For the first problem, they derived an integer programming based upper bound on the minimum distance for LRCs, and then gave an explicit bound by solving the integer programming problem. The explicit bound applies all LRCs satisfying  $n_1 > n_2$ , where  $n_1 = \lceil \frac{n}{r+1} \rceil$  and  $n_2 = n_1(r+1) - n$ . For the second problem, they presented a construction of linear LRCs that attains the explicit bound for  $n_1 > n_2$ . Therefore, they had completely solved the two problems under the condition  $n_1 > n_2$ . A similar result can be found in [44] using matroid theory.

In this paper, we first deal with the open case remained in [40] and derive an improved upper bound for the minimum distances of LRCs. We also give an explicit construction for LRCs attaining this bound.

There are lots of other works devoted to the locality in the handling of multiple node failures, such as [31, 36, 37], considering LRCs which permit parallel access of “hot data”, the works of [26, 41] studying LRCs with general local repair groups, and the work [30] which proposed sequential local repair. Very recently, Wang et al. [42] proposed a binary LRC construction achieving any locality and availability with very high code rate. An LRC code  $C[n, k, d]$  is said to have locality  $r$  and availability  $t$ , if for any codeword  $y \in C$ , any symbol  $y_i$  of  $y$  can be computed from some other  $r$  symbols of  $y$ , and furthermore there are  $t$  disjoint ways to reconstruct  $y_i$ . Unfortunately, the minimum distance of the codes constructed in [42] is too small, saying  $t + 1$ .

The second part of this paper deals with constructions of binary LRCs  $C[n, k, d]$  with any locality  $r$  and availability 2 which have both high code rate and large minimum distance. We first give a graphical model for binary LRCs. We then use graphs with long girth to give a high rate code construction. Comparing with the constructions [30, 42], our codes have a slight decline of rate, however, our codes have much larger minimum distance ( $d = O(\log n)$ ).

This paper is organized as follows. Section 2 reviews some elementary results that will be used in this paper. Section 3 solves the integer programming problem put forward in [40], and gives an explicit upper bound for LRCs satisfying  $n_1 \leq n_2$ . Then Section 4 presents an explicit construction attaining this bound. Section 5 gives a construction of a family of LRCs with any locality  $r$  and availability 2 having code rate  $\frac{r-1}{r+1}$  and minimum distance  $O(\log n)$  where  $n$  is the length of the code. Finally, Section 6 concludes the paper.

## 2. PRELIMINARIES

In [40], the authors derived an integer programming based bound on the minimum distance of any LRC. Define

$$(2.1) \quad \Psi(x) = \max_{\substack{s, t_1, \dots, t_s \\ a_1, \dots, a_s}} \min_{l, h_1, \dots, h_l} (xr + 1 - \sum_{i=1}^{l-1} (a_{h_i} - t_{h_i})), \forall 1 \leq x \leq n_1,$$

where  $s, t_1, \dots, t_s, a_1, \dots, a_s$  satisfy

$$\begin{cases} t_1 + \dots + t_s = n_1; \\ a_1 + \dots + a_s = n_2; \\ a_i \geq t_i - 1, \quad \forall 1 \leq i \leq s; \\ s \geq 1; \quad t_i \geq 1, \quad \forall 1 \leq i \leq s. \end{cases}$$

and  $l, h_1, \dots, h_l$  satisfy

$$(2.2) \quad t_{h_1} + \dots + t_{h_{l-1}} < x \leq t_{h_1} + \dots + t_{h_l}.$$

**Theorem 2.1** ([40]). *For any  $[n, k, d]$  LRC,*

$$d \leq n - k + 1 - \eta,$$

where  $\eta = \max\{x : \Psi(x) - x < k\}$ .

Next, we review the construction of Tamo and Barg [37] as their construction gives some optimal codes for the bound we will obtain later. Furthermore, we will employ their construction to get more optimal codes meeting our bound.

Let  $A \subset F$ , and let  $\mathcal{A}$  be a partition of  $A$  into  $m$  subsets  $A_i$ . Consider the set of polynomials  $F_{\mathcal{A}}[x]$  of degree less than  $|A|$  that are constant on the blocks of the partition:

$$F_{\mathcal{A}}[x] = \{f \in F[x] : f \text{ is constant on } A_i, i = 1, \dots, m; \deg f < |A|\}.$$

The annihilator of  $A$  is the smallest-degree monic polynomial  $h_A$  such that  $h_A(a) = 0$  if  $a \in A$ , i.e.,  $h_A(x) = \prod_{a \in A} (x - a)$ . Observe that the set  $F_{\mathcal{A}}[x]$  with the usual addition and multiplication modulo  $h(x)$  becomes a commutative algebra with identity. Since the polynomials  $F_{\mathcal{A}}[x]$  are constant on the elements of  $\mathcal{A}$ , we write  $f(A_i)$  to refer to the value of the polynomial  $f$  on the set  $A_i \in \mathcal{A}$ .

**Proposition 2.2** ([37]). *Let  $\alpha_1, \dots, \alpha_m$  be distinct nonzero elements of  $F$ , and let  $g$  be the polynomial of degree  $\deg(g) < |A|$  that satisfies  $g(A_i) = \alpha_i$  for all  $i = 1, \dots, m$ , i.e.,*

$$g(x) = \sum_{i=1}^m \alpha_i \sum_{a \in A_i} \prod_{b \in A \setminus a} \frac{x - b}{a - b}.$$

*Then the polynomials  $1, g, \dots, g^{m-1}$  form a basis of  $F_{\mathcal{A}}[x]$ .*

**Proposition 2.3** ([37]). *There exist  $m$  integers  $0 = d_0 < d_1 < \dots < d_{m-1} < |A|$  such that the degree of each polynomial in  $F_A[x]$  is  $d_i$  for some  $i$ .*

**Corollary 2.4** ([37]). *Assume that  $d_1 = r + 1$ , namely there exists a polynomial  $g$  in  $F_A[x]$  of degree  $r + 1$ , then  $d_i = i(r + 1)$  for all  $i = 0, \dots, m - 1$ , and the polynomials  $1, g, \dots, g^{m-1}$  defined in Proposition 2.2, form a basis for  $F_A[x]$ .*

**Construction 2.5** ([37]). 1. *Let  $F$  be a finite field, and let  $A \subset F$  be a subset such that  $|A| = n$ ,  $n \bmod (r + 1) = s \neq 0, 1$ . Assume also that  $k + 1$  is divisible by  $r$  (this assumption is nonessential).*

2. *Let  $\mathcal{A}$  be a partition of  $A$  into  $m$  subsets  $A_1, \dots, A_m$  such that  $|A_i| = r + 1, 1 \leq i \leq m - 1$  and  $1 < |A_m| = s < r + 1$ . Let  $g(x)$  be a polynomial of degree  $r + 1$ , such that its powers  $1, g, \dots, g^{m-1}$  span the algebra  $F_A[x]$ . W.L.O.G., assume that  $g$  vanishes on the set  $A_m$ , otherwise one can take the powers of the polynomial  $g(x) - g(A_m)$  as the basis for the algebra.*

3. *Let  $a = (a_0, \dots, a_{r-1}) \in F^k$  be the input information vector, such that each  $a_i$  for  $i \neq s - 1$  is a vector of length  $\frac{k+1}{r}$  and  $a_{s-1}$  is of length  $\frac{k+1}{r} - 1$ . Define the encoding polynomial*

$$f_a(x) = \sum_{i=0}^{s-2} \sum_{j=0}^{\frac{k+1}{r}-1} a_{i,j} g(x)^j x^i + \sum_{j=1}^{\frac{k+1}{r}-1} a_{s-1,j} g(x)^j x^{s-1} + \sum_{i=s}^{r-1} \sum_{j=0}^{\frac{k+1}{r}-1} a_{i,j} g^j(x) x^{i-s} h_{A_m}(x).$$

*The code is defined as the set of evaluations of  $f_a(x)$ ,  $a \in F^k$ .*

**Theorem 2.6** ([37]). *The code given by Construction 2.5 is an  $[n, k, r]$  LRC code with minimum distance satisfying*

$$d \geq n - k - \left\lceil \frac{k}{r} \right\rceil + 1.$$

### 3. UPPER BOUNDS OF THE MINIMUM DISTANCE

In this section, we solve the integer programming problem (2.1), and derive an explicit upper bound for all LRCs satisfying  $n_1 \leq n_2$ . Then we make comparisons with the bound (1.1) to show the improvements of our explicit bound. Actually, in the next section we will show our bound is tight for the case  $n_1 \leq n_2$ .

**Theorem 3.1.** *For  $1 \leq x \leq n_1$  and  $n_1 \leq n_2$ ,*

$$\Psi(x) = xr + 1.$$

*Proof.* 1. Set

$$\begin{cases} s &= 1, \\ t_1 &= n_1, \\ a_1 &= n_2. \end{cases}$$

Then we have

$$\Psi(x) \geq \min_{l, h_1, \dots, h_l} (xr + 1 - \sum_{i=1}^{l-1} (a_{h_i} - t_{h_i})) = xr + 1.$$

2. Assume that for some  $1 \leq x \leq n_1$ ,

$$\Psi(x) \geq xr + 2.$$

Then there exist integers  $s$  and  $t_i, a_i, 1 \leq i \leq s$ , satisfying the constraints of the integer programming and

$$\min_{l, h_1, \dots, h_l} (xr + 1 - \sum_{i=1}^{l-1} (a_{h_i} - t_{h_i})) \geq xr + 2.$$

Therefore for all integers  $l$  and  $h_1, \dots, h_l \in [s]$  satisfying the constraint (2.2), we have

$$(3.1) \quad \sum_{i=1}^{l-1} (a_{h_i} - t_{h_i}) \leq -1.$$

If there is some  $i$  such that  $t_i \geq x$ , let  $h_1 = i$  in the constraint (2.2), then

$$\sum_{i=1}^{l-1} (a_{h_i} - t_{h_i}) = 0,$$

which contradicts to (3.1). And hence, the assumption  $\Psi(x) \geq xr + 2$  does not hold, and we finish the proof.

So we assume  $t_i < x, \forall 1 \leq i \leq s$ . For  $1 \leq i \leq s$ , define

$$b_i = a_i - t_i.$$

W.L.O.G, we assume that  $b_1 \leq b_2 \leq \dots \leq b_s$ . Since  $t_i < x$ , we can find  $i_0 = 1, i_1, i_2, \dots, i_p < s$  satisfying that

$$\begin{aligned} t_1 + \dots + t_{i_1-1} &< x \leq t_1 + \dots + t_{i_1}, \\ t_{i_1} + \dots + t_{i_2-1} &< x \leq t_{i_1} + \dots + t_{i_2}, \\ &\dots \\ t_{i_{p-1}} + \dots + t_{i_p-1} &< x \leq t_{i_{p-1}} + \dots + t_{i_p}, \\ t_{i_p} + \dots + t_s &< x. \end{aligned}$$

Then we have

$$\begin{aligned} \sum_{i=1}^{i_1-1} b_i &\leq -1, \\ \sum_{i=i_1}^{i_2-1} b_i &\leq -1, \\ &\dots, \\ \sum_{i=i_{p-1}}^{i_p-1} b_i &\leq -1. \end{aligned}$$

Noting  $\sum_{i=1}^s b_i = n_2 - n_1 \geq 0$ , we get  $\sum_{i=i_p}^s b_i > 0$ . Because  $x \leq n_1$ , then  $p \geq 1$ , we can consider the last two parts of  $t_i$ 's in the reverse order  $t_s, \dots, t_{i_p}, \dots, t_{i_{p-1}}$ . From the definition of  $i_p$ , we know  $\sum_{m=s}^{i_{p-1}} t_m \geq x$  and  $\sum_{m=s}^{i_{p-1}} b_m \geq 0$ . So there exists  $q$  satisfying  $\sum_{m=s}^{i_{q-1}} t_m < x \leq \sum_{m=s}^{i_q} t_m$  and  $\sum_{m=s}^{i_{q-1}} b_m \leq -1$ . On the other hand  $b_{i_{p-1}} \leq b_{i_p} \leq \dots \leq b_s$ , we get  $b_{i_{p-1}}, \dots, b_{i_q} < 0$ , which contradicts to  $\sum_{m=s}^{i_{p-1}} b_m \geq 0$ .

Thus we have  $\Psi(x) \leq xr + 1$ . □

**Theorem 3.2.** For any  $[n, k, d]$  LRC with  $n_1 \leq n_2$ , where  $n_1 = \lceil \frac{n}{r+1} \rceil$  and  $n_2 = n_1(r+1) - n$ , it holds that

$$(3.2) \quad d \leq n - k + 1 - (\lceil \frac{k-1}{r-1} \rceil - 1).$$

*Proof.* This follows from Theorems 2.1 and 3.1.  $\square$

Since the bound (3.2) in Theorem 3.2 holds for  $n_1 \leq n_2$ , all the comparisons we make below are under the condition  $n_1 \leq n_2$ .

The bound (1.1) given by Gopalan et al. [13] is the first upper bound on the minimum distance of LRCs. Since  $r < k$  (the natural condition that LRCs require), we always have  $\lceil \frac{k-1}{r-1} \rceil \geq \lceil \frac{k}{r} \rceil$ . So the bound (3.2) generally provides a tighter upper bound than the bound (1.1).

Specially, we assume  $k = ur + v$  for some integers  $u, v$  and  $0 \leq v \leq r - 1$ , then

$$d \leq n - k + 1 - (\lceil \frac{k-1}{r-1} \rceil - 1) = \begin{cases} n - k - u + 1, & u + v \leq r; \\ u - k - u, & u + v > r. \end{cases}$$

#### 4. CODE CONSTRUCTION WHEN $n_1 \leq n_2$

In this section, we present an explicit construction of LRCs attaining the bound (3.2) in some cases. The idea of construction comes from [37].

**Theorem 4.1.** When  $n_1 \leq n_2$  and  $u + v > r$ ,  $n_2 \neq r$ , the bound (3.2) is achievable.

*Proof.* This follows from Theorems 2.6 and 3.1.  $\square$

Modifying the construction in the above theorem, we show that the bound (3.2) is also tight in other cases.

**Construction 4.2.** Let  $F$  be a finite field, and let  $A \subset F$  be a subset such that  $|A| = n$ .

- (1) Since  $n = n_1(r+1) - n_2 = n_1r - (n_2 - n_1)$ , let  $\mathcal{A}$  be a partition of  $A$  into  $n_1$  subsets  $A_1, \dots, A_{n_1}$  such that  $|A_i| = r, 1 \leq i \leq n_1 - 1$  and  $|A_{n_1}| = s = r - (n_2 - n_1) > 1$ . Let  $g(x)$  be a polynomial of degree  $r$ , such that its powers  $1, g, \dots, g^{n_1-1}$  span the algebra  $F_{\mathcal{A}}[x]$ . W.L.O.G., we assume that  $g$  vanishes on the set  $A_{n_1}$  and  $u + v = s$  (this assumption is nonessential, in fact we only need  $u + v \leq s$ ).
- (2) Let  $a = (a_0, \dots, a_{r-2}) \in F^k$  be the input information vector, such that  $a_i$  is a vector of length  $u + 1$  for  $0 \leq i \leq s - 1$  and  $a_i$  is a vector of length  $u$  for  $i \geq s$ . Define the encoding polynomial

$$f_a(x) = \sum_{i=0}^{s-1} \sum_{j=0}^u a_{i,j} g(x)^j x^i + \sum_{i=s}^{r-1} \sum_{j=0}^{u-1} a_{i,j} g(x)^j x^{i-s} h_{A_{n_1}}(x),$$

where  $h_{A_{n_1}} = \prod_{a \in A_{n_1}} (x - a)$ .

The code is defined as the set of evaluations of  $f_a(x)$ ,  $a \in F^k$ .

**Theorem 4.3.** Keep the notation as above. The code given in Construction 4.2 is an  $[n, k]$  LRC with locality  $r - 1$  and minimum distance

$$d \geq n - k - u + 1.$$

*Proof.* Since the encoding is linear and the encoding polynomials have degree at most

$$\max\{ur + s - 1, (u - 1)r + r - 1\} = ur + u + v - 1 = k + u - 1$$

we have  $d \geq n - \deg(f) \geq n - k - u + 1$ . The locality property is similar to Construction (2.5). If the erased symbol  $f_a(x)$  lies in  $x \in A_{n_1}$ , by interpolating the other  $s - 1$  points in  $A_{n_1}$  we get a polynomial of degree at most  $s - 2$  to recover  $f_a(x)$ . Otherwise, we use  $r - 1$  interpolation points to find a polynomial of degree at most  $r - 2$  to recover  $f_a(x)$ . So, it is an LRC code with locality  $(r - 1)$ . The result follows.  $\square$

As a corollary, we obtain more tight range for the bound (3.2).

**Corollary 4.1.** *When  $n_1 < n_2$  and  $u + v + n_2 - n_1 \leq r$ , the bound (3.2) is achievable.*

*Proof.* We can view the code constructed above as an LRC with locality  $r$ , then the corollary follows directly from Theorems 3.2 and 4.3.  $\square$

## 5. GRAPH-BASED CONSTRUCTION OF LRCs WITH ARBITRARY LOCALITY AND AVAILABILITY 2

Very recently, Wang et al. [42] proposed a binary LRC construction achieving any locality and availability with very high code rate. In this section, we first give a graphical model for binary LRCs. Secondly, we consider the special case  $t = 2$ , i.e., there are two disjoint repair ways for any coordinate. We give a high rate code construction. Comparing with the construction [42], our codes have a slight decline of rate, however, our codes have much larger minimum distances.

Recall that an LRC  $C[n, k, d]$  with locality  $r$  and availability  $t$  satisfies the following property: for any codeword  $y \in C$ , any symbol  $y_i$  of  $y$  can be computed from some other  $r$  symbols of  $y$ , and furthermore there are  $t$  disjoint ways to reconstruct  $y_i$ .

**Proposition 5.1** ([36]). *For a linear code  $C[n, k, d]$  with locality  $r$  and availability  $t$ , the rate of the code satisfies*

$$\frac{k}{n} \leq \prod_{i=1}^t \frac{1}{1 + \frac{1}{ir}}.$$

The bound in the above proposition can not be achieved in most cases. Wang et al. [42] gave a construction from the incidence matrices of some combinatorial designs:

**Proposition 5.2** ([42]). *For any  $r$  and  $t$ , there are binary linear codes  $C[n, k, d]$  with locality  $r$  and availability  $t$  satisfying*

$$\frac{k}{n} = \frac{r}{r + t} \quad \text{and} \quad d = t + 1.$$

Note that for fixed  $t$ , the minimum distance of the construction in the above proposition is fixed. But as discussed at the beginning of this paper, the minimum distance of LRCs is a very important metric, especially for multiple erasures. So how to construct LRCs with high rate, large minimum distance, any locality and availability is the issue we care in the following content.

Next, we only consider the binary case. The method works as well for non-binary cases. To construct a binary LRC  $C[n, k, d]$  with locality  $r$  and availability  $t$ , it is equivalent to construct a parity check matrix  $H$  such that each column has Hamming weight  $\geq t$  and each

row has Hamming weight  $\leq r + 1$  such that the inner product of any two rows is 1. Note that rows of  $H$  might be linearly dependent. Corresponding to this parity check matrix

$$H = (h_{i,j})_{1 \leq i \leq m, 1 \leq j \leq n},$$

there is a bipartite graph  $G$  whose bi-adjacent matrix is  $H$ . Explicitly, the graph  $G = (V, E)$  is defined as following:

- The set  $V$  of vertices is separated into two parts  $\{c_1, c_2, \dots, c_m\}$  and  $\{x_1, x_2, \dots, x_n\}$  which represent rows and columns, respectively. The vertices  $\{c_1, c_2, \dots, c_m\}$  are always called constraints, and the vertices  $\{x_1, x_2, \dots, x_n\}$  are called variables.
- The set  $E$  of edges: there are no edges connecting vertices in the same part, all the edges are connecting vertices from the distinct parts. Precisely, there is an edge between  $c_i$  and  $x_j$  if and only if  $h_{i,j} = 1$  for any  $1 \leq i \leq m, 1 \leq j \leq n$ .

**Example 5.1.** For the matrix

$$H = \begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \end{pmatrix},$$

the corresponding graph is Figure 1. The matrix  $H$  defines a  $[9, 4]$  binary code with locality 2 and availability 2.

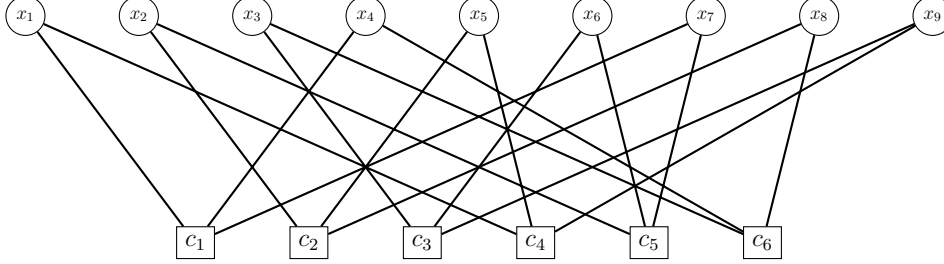


FIGURE 1. The bipartite graph  $G$  representing  $H$

In the literature, the bipartite graph is called Tanner graph. Recall that the degree of a vertex  $v \in V$  is defined to be the number of edges connecting the vertex, and denoted by  $\deg(v)$ . In this setting, in order to construct a binary LRC  $C[n, k, d]$  with locality  $r$  and availability  $t$ , the problem is reduced to constructing a bipartite graph  $G$  with vertices  $\{c_1, c_2, \dots, c_m\} \cup \{x_1, x_2, \dots, x_n\}$  such that  $\deg(c_i) \leq r + 1$  and  $\deg(x_j) \geq t$  for any  $1 \leq i \leq m, 1 \leq j \leq n$ . Meanwhile, to simplify the discussion, we only consider the regular case. That is, the Tanner graph is a regular graph, where

$$\deg(c_1) = \deg(c_2) = \dots = \deg(c_m) = r + 1$$

and

$$\deg(x_1) = \deg(x_2) = \dots = \deg(x_n) = t.$$

In this case, we have the following lower bound for the corresponding code:

**Proposition 5.3.** *The rate  $\rho$  of the binary code is*

$$\rho \geq 1 - \frac{t}{r+1}.$$

*Proof.* By counting the number of 1's in the bi-adjacent matrix  $H$  of  $G$ , we have

$$m(r+1) = nt, \quad \text{or} \quad \frac{m}{n} = \frac{t}{r+1}.$$

So the rate of the code is

$$\rho = 1 - \frac{\text{Rank}(H)}{n} \geq 1 - \frac{m}{n} = 1 - \frac{t}{r+1},$$

where  $\text{Rank}(H)$  is the  $\mathbb{F}_2$ -rank of  $H$ . □

It is a very tough work to compute the exact value of  $\text{Rank}(H)$  in general. But it is an important issue in many application scenarios. For graphs with strong combinatorial property, computing  $\text{Rank}(H)$  attracts lots of interests [4, 5, 8, 18, 21, 34].

There are advantages of graphical representation of codes [10, 19]. It generalizes low-density parity-check codes, convolutional codes, trellis codes, classical linear system theory, behavior systems theory, etc. Fast algorithms on graphs give efficient encoding and decoding algorithms, such as the sum-product algorithm, BCJR algorithm, Viterbi algorithm, etc. In our specific case of LRCs, when the information of any node is not available or damaged, it is easy to recover the information by adding the information of neighboring variable vertices of any neighbor of the node in the graph. Even if many variable nodes are damaged, we can track in the graph for the intact information to recover the damaged nodes provided that the number of damaged nodes is less than the minimum distance of the code. This is our motivation to enlarge the minimum distance of the LRCs with the required locality and availability as large as possible.

Next, we restrict ourselves to the case  $t = 2$  where there are two disjoint repair options for each coordinate. In other word, the degree of  $x_j$  ( $1 \leq j \leq n$ ) is two in the Tanner graph. In this case, the rate of the code is

$$\geq \frac{r-1}{r+1}$$

which might be smaller than that of [42] by difference (at most)

$$\frac{r}{r+2} - \frac{r-1}{r+1} = \frac{2}{(r+1)(r+2)}.$$

By slight sacrifice of the code rate, we can construct codes with much larger minimum distance. More concretely, the code in [42] has minimum distance 3, but our codes have minimum distance  $O(\log n)$ .

Since  $\deg(x_j) = 2$  for all  $1 \leq j \leq n$ , the Tanner graph  $G$  can be reduced to a smaller graph  $G_{red}$ :

- The vertices are  $\{c_1, c_2, \dots, c_m\}$ .
- There is an edge between  $c_i$  and  $c_j$  if and only if  $c_i$  and  $c_j$  connect some  $x_l$  simultaneously in the graph  $G$ .

The reduced graph  $G_{red}$  is an  $(r+1)$ -regular graph. One could also refer the reduced graph  $G_{red}$  as another graphical model of the code  $C$ . The difference between the two graphs is

that Tanner graph considers constraints (or rows) as one part of the bipartite graph, but the reduced graph considers the constraints as the edges of the graph.

**Example 5.2.** *Continue Example 5.1. The reduced graph  $G_{red}$  is Figure 2.*

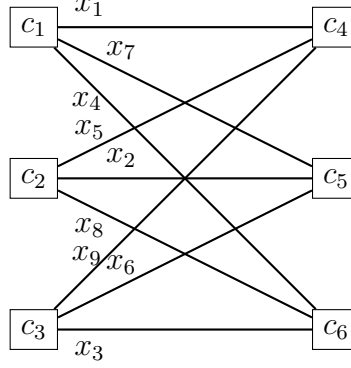


FIGURE 2. The reduced graph representing  $H$

To analyze the minimum distance of the code, we need one more index of the Tanner graph or the reduced graph. The girth of a graph is the length of a shortest cycle in the graph. Since only graphs without multi-edges are involved in this paper, it is easy to see that the girth of a graph is 0 or  $\geq 3$ , and the girth of a bipartite is an even integer: 0 or  $\geq 4$ .

**Theorem 5.1.** *Let  $G_{red}$  be an  $(r + 1)$ -regular graph with  $m$  vertices and girth  $g$ . Extend the graph  $G_{red}$  to a bipartite graph  $G$  with regularity 2 and  $r + 1$ . The null space of the bi-adjacent matrix  $H$  of  $G$  defines our binary linear code  $C$ . Then the code  $C$  has length  $\frac{m(r+1)}{2}$ , dimension  $\geq \frac{m(r-1)}{2} + 1$ , minimum distance  $g$ , locality  $r$ , and availability 2.*

*Proof.* The only thing we need to prove is that minimum distance  $d = g$ . On one hand, W.L.O.G., let  $c_1, c_2, \dots, c_g$  be a cycle of length  $g$  in  $G_{red}$ . Then it is extended to a cycle of length  $2g$  in  $G$ , saying  $c_1, x_1, c_2, x_2, \dots, c_g, x_g$ . By the fact  $\deg(x_j) = 2$ , the restriction of the parity check matrix  $H$  to the columns  $x_1, x_2, \dots, x_g$  is

	$x_1$	$x_2$	$x_3$	$\dots$	$x_g$
$c_1$	1	1	0	$\dots$	0
$c_2$	0	1	1	$\dots$	0
$c_3$	0	0	1	$\dots$	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$c_g$	1	0	0	$\dots$	1
$\vdots$	0	0	0	$\vdots$	0

which defines a codeword with support set  $\{1, 2, \dots, g\}$ . So the minimum distance  $d \leq g$ .

On the other hand, let  $c$  be a codeword of Hamming weight  $d$ . W.L.O.G., assume the support of  $c$  is  $\{1, 2, \dots, d\}$ . By non-zero location chasing, we prove  $d \geq g$ . We can also assume the variables  $x_1, x_2$  are connected to the constraint  $c_1$ . Now,  $x_2$  has the other neighboring constraint, saying  $c_2$ . As the codeword  $c$  must satisfy the constraint  $c_2$ , there is at least one  $x_j$  connecting  $c_2$  for some  $1 \leq j \leq d, j \neq 2$ . If  $j = 1$  then we get a cycle of length 4 in  $G$ , so

$g = 2 \leq d$  and the proof is finished. Otherwise, assume  $x_3$  connects  $c_2$ . If  $x_3$  also connects  $c_1$ , then the proof is finished as the same as the previous. Otherwise,  $x_3$  connects the other constraint. Then iterate the same procedure. One can finally get a cycle of length  $\leq 2d$  in the graph  $G$ . So the girth of the reduced graph  $G_{red}$  is  $g \leq d$ .

In conclusion, we have proved the minimum distance  $d = g$ . □

The theorem extends the result of [14, Proposition 2]. For non-binary case, Chen et al. [6] proposed how to enlarge the minimum distance by choosing proper non-zero elements at the non-zero locations of  $H$ . By Theorem 5.1, in order to construct a binary LRC with rate  $\frac{r-1}{r+1}$ , locality  $r$ , availability 2, and minimum distance as large as possible, we need to construct an  $(r+1)$ -regular graph with girth  $g$  as large as possible. This latter problem of graph construction has been extensively studied in extremal graph theory.

Let  $g(m, r)$  denote the largest possible girth of an  $(r+1)$ -regular graph of size at most  $m$ , then for fixed  $r$  and asymptotically growing  $m$  we have

$$(5.1) \quad \left(\frac{4}{3} - o(1)\right) \log_r m \leq g(m, r) \leq (2 + o(1)) \log_r m.$$

The second inequality in (5.1) is a version of the Moore bound [3, Theorem III.1]. Note that the Moore bound is not achievable in most cases. The girths of random Cayley graphs are tested in [12]. The first explicit construction can be found in [22] for graphs with degree 4 and large girth  $\geq 0.83 \log_r m$  and those with arbitrary degree and large girth  $\geq 0.44 \log_r m$ , the latter of which was later improved in [17] to  $\geq 0.48 \log_r m$ . Erdős and Sachs [7] described a simple procedure yielding families of graphs with large girth  $\log_r m$ . Examples of graphs with arbitrary degree and large girth  $\geq \frac{4}{3} \log_r m$  are given in [2, 20, 23, 24, 43]. Using these explicit constructions, we can obtain

**Theorem 5.2.** *Let  $G_{red}$  be an  $(r+1)$ -regular graph with  $n$  edges and girth  $g = O(\log n)$ . Extend the graph  $G_{red}$  to a bipartite graph  $G$  with regularity 2 and  $r+1$ . The null space of the bi-adjacent matrix  $H$  of  $G$  defines our binary linear code  $C$ . Then the code  $C$  has length  $n$ , dimension  $\geq \frac{n(r-1)}{r+1} + 1$ , minimum distance  $O(\log n)$ , locality  $r$ , and availability 2.*

Comparing with the constructions in [30, 42], our codes have a slight decline of rate, however, our codes have much larger minimum distances. Comparing with the construction of [37, Theorem 4.1], the minimum distances of their codes are very large apparently. On one hand, their construction relies highly on the size of the finite field, so their method can not be employed for the binary case. On the other hand, if their code rate achieves  $\frac{r-1}{r+1}$ , the minimum distance of their code degenerates to 1. Actually, the minimum distance  $O(\log n)$  in the above theorem is already optimal in the case  $t = 2$  by [11, Theorem 2.5].

**Remark 5.1.** *Analogously to the performance of random linear codes, for general locality  $r$  and availability  $t \geq 3$ , the codes constructed from random  $(r+1, t)$ -regular bipartite graphs have minimum distances with growing rate linearly to the length of the code with very high probability [11, Theorem 2.4]. Within our knowledge, there is no deterministic construction for  $(r+1, t)$ -regular bipartite graphs (arbitrary  $r$  and  $t$ ) such that the corresponding codes have non-zero relative minimum distance  $\frac{d}{n}$  asymptotically.*

## 6. CONCLUSIONS

In the first part of this paper we studied the open problem in [40]: when  $n_1 \leq n_2$ , what is the largest possible minimum distance for an  $[n, k]$  LRC? How to construct an  $[n, k]$  LRC with the largest possible minimum distance? For the first problem, we solve the linear integer programming in the case  $n_1 \leq n_2$  and derive a new upper bound which is always better than the classic bound (1.1). For the second problem, we find out that the construction of Tamo and Barg [37] is actually optimal when  $n_1 \leq n_2$  and  $u + v > r, v \neq r$ . Using another interpolation polynomial, we present a construction of optimal LRCs when  $n_1 < n_2$  and  $u + v + n_2 - n_1 \leq r$ .

In the second part of this paper, we presented a graphical model for binary LRC with any locality and any availability. In particular, for any locality and availability 2, we use the deep results from extremal graph theory to give a code construction which produces good LRCs in the sense that these codes satisfy the locality and availability request and they have high code rates and large (indeed optimal) minimum distances.

## REFERENCES

- [1] A. Barg, I. Tamo, and S.G. Vladut. Locally recoverable codes on algebraic curves. *CoRR*, abs/1501.04904, 2015.
- [2] N.L. Biggs and M.J. Hoare. The sextet construction for cubic graphs. *Combinatorica*, 3(2):153–165, 1983.
- [3] B. Bollobás. *Extremal Graph Theory*. Dover Books on Mathematics. Dover Publications, 2004.
- [4] A.E. Brouwer and C.A. Van Eijl. On the  $p$ -rank of the adjacency matrices of strongly regular graphs. *Journal of Algebraic Combinatorics*, 1(4):329–346, 1992.
- [5] D.B. Chandler and Q. Xiang. Cyclic relative difference sets and their  $p$ -ranks. *Designs, Codes and Cryptography*, 30(3):325–343, 2003.
- [6] C. Chen, B. Bai, G. Shi, X. Wang, and X. Jiao. Nonbinary LDPC codes on cages: Structural property and code optimization. *Communications, IEEE Transactions on*, 63(2):364–375, Feb 2015.
- [7] P. Erdős and H. Sachs. Reguläre graphen gegebener taillenweite mit minimaler knotenzahl. *Wiss. Z. Martin-Luther-Univ. Halle-Wittenberg, Math.-Naturwiss. Reihe*, 12:251–258, 1963.
- [8] R. Evans, H.D.L. Hollmann, C. Krattenthaler, and Q. Xiang. Gauss sums, jacobi sums, and  $p$ -ranks of cyclic difference sets. *Journal of Combinatorial Theory, Series A*, 87(1):74–119, 1999.
- [9] M. Forbes and S. Yekhanin. On the locality of codeword symbols in non-linear codes. *Discrete Mathematics*, 324:78–84, 2014.
- [10] G.D. Jr. Forney. Codes on graphs: normal realizations. *Information Theory, IEEE Transactions on*, 47(2):520–548, Feb 2001.
- [11] R.G. Gallager. *Low-Density Parity-Check Codes*. Cambridge, MA: MIT Press, 1963.
- [12] A. Gamburd, S. Hoory, M. Shahshahani, A. Shalev, and B. Virag. On the girth of random Cayley graphs. *Random Struct. Algorithms*, 35(1):100–117, 2009.
- [13] P. Gopalan, C. Huang, H. Simitci, and S. Yekhanin. On the locality of codeword symbols. *Information Theory, IEEE Transactions on*, 58(11):6925–6934, 2012.
- [14] X.-Y. Hu, M.P.C. Fossorier, and E. Eleftheriou. On the computation of the minimum distance of low-density parity-check codes. In *Communications, 2004 IEEE International Conference on*, volume 2, pages 767–771 Vol.2, June 2004.
- [15] C. Huang, M. Chen, and J. Li. Pyramid codes: Flexible schemes to trade space for access efficiency in reliable data storage systems. In *Network Computing and Applications, 2007. NCA 2007. Sixth IEEE International Symposium on*, pages 79–86, 2007.
- [16] C. Huang, H. Simitci, Y. Xu, A. Ogus, B. Calder, P. Gopalan, J. Li, S. Yekhanin, et al. Erasure coding in windows azure storage. In *USENIX Annual Technical Conference*, pages 15–26, 2012.
- [17] W. Imrich. Explicit construction of regular graphs without small cycles. *Combinatorica*, 4(1):53–59, 1984.

- [18] S.J. Johnson and S.R. Weller. Codes for iterative decoding from partial geometries. *Communications, IEEE Transactions on*, 52(2):236–243, Feb 2004.
- [19] F.R. Kschischang, B.J. Frey, and H.-A. Loeliger. Factor graphs and the sum-product algorithm. *Information Theory, IEEE Transactions on*, 47:498–519, 1998.
- [20] A. Lubotzky, R. Phillips, and P. Sarnak. Ramanujan graphs. *Combinatorica*, 8(3):261–277, 1988.
- [21] F.J. MacWilliams and H.B. Mann. On the  $p$ -rank of the design matrix of a difference set. *Information and Control*, 12(5):474–488, 1968.
- [22] G.A. Margulis. Explicit constructions of graphs without short cycles and low density codes. *Combinatorica*, 2(1):71–78, 1982.
- [23] G.A. Margulis. Explicit group-theoretical constructions of combinatorial schemes and their application to the design of expanders and concentrators. *Probl. Peredachi Inf.*, 24(1):51–60, 1988.
- [24] M. Morgenstern. Existence and explicit constructions of  $q+1$  regular Ramanujan graphs for every prime power  $q$ . *Journal of Combinatorial Theory, Series B*, 62(1):44–62, 1994.
- [25] F. Oggier and A. Datta. Self-repairing homomorphic codes for distributed storage systems. In *INFOCOM, 2011 Proceedings IEEE*, pages 1215–1223, 2011.
- [26] L. Parnes-Juarez, H.D.L. Hollmann, and F. Oggier. Locally repairable codes with multiple repair alternatives. In *Information Theory Proceedings (ISIT), 2013 IEEE International Symposium on*, pages 892–896, 2013.
- [27] D.S. Papailiopoulos and A.G. Dimakis. Locally repairable codes. In *information theory proceedings (ISIT), 2012 IEEE international symposium on*, pages 2771–2775, 2012.
- [28] D.S. Papailiopoulos, J. Luo, A.G. Dimakis, C. Huang, and J. Li. Simple regenerating codes: Network coding for cloud storage. In *INFOCOM, 2012 Proceedings IEEE*, pages 2801–2805, 2012.
- [29] N. Prakash, G.M. Kamath, V. Lalitha, and P.V. Kumar. Optimal linear codes with a local-error-correction property. In *Information Theory Proceedings (ISIT), 2012 IEEE International Symposium on*, pages 2776–2780, 2012.
- [30] N. Prakash, V. Lalitha, and P.V. Kumar. Codes with locality for two erasures. In *Information Theory (ISIT), 2014 IEEE International Symposium on*, pages 1962–1966, June 2014.
- [31] A.S. Rawat, D.S. Papailiopoulos, A.G. Dimakis, and S. Vishwanath. Locality and availability in distributed storage. *arXiv preprint arXiv:1402.2011*, 2014.
- [32] M. Sathiamoorthy, M. Asteris, D. Papailiopoulos, A.G. Dimakis, R. Vadali, S. Chen, and D. Borthakur. XORing elephants: Novel erasure codes for big data. In *Proceedings of the VLDB Endowment*, volume 6, pages 325–336. VLDB Endowment, 2013.
- [33] N. Silberstein, A.S. Rawat, O. Koyluoglu, and S. Vishwanath. Optimal locally repairable codes via rank-metric codes. In *Information Theory Proceedings (ISIT), 2013 IEEE International Symposium on*, pages 1819–1823, 2013.
- [34] K.J.C. Smith. On the  $p$ -rank of the incidence matrix of points and hyperplanes in a finite projective geometry. *Journal of Combinatorial Theory*, 7(2):122–129, 1969.
- [35] W. Song, S.H. Dau, C. Yuen, and J. Li. Optimal locally repairable linear codes. *Selected Areas in Communications, IEEE Journal on*, 32(5):1019–1036, 2014.
- [36] I. Tamo and A. Barg. Bounds on locally recoverable codes with multiple recovering sets. In *2014 IEEE International Symposium on Information Theory, Honolulu, HI, USA, June 29-July 4, 2014*, pages 691–695, 2014.
- [37] I. Tamo and A. Barg. A family of optimal locally recoverable codes. *Information Theory, IEEE Transactions on*, 60(8):4661–4676, 2014.
- [38] I. Tamo, A. Barg, S. Goparaju, and A.R. Calderbank. Cyclic LRC codes and their subfield subcodes. *CoRR*, abs/1502.01414, 2015.
- [39] I. Tamo, D.S. Papailiopoulos, and A.G. Dimakis. Optimal locally repairable codes and connections to matroid theory. In *Information Theory Proceedings (ISIT), 2013 IEEE International Symposium on*, pages 1814–1818, 2013.
- [40] A. Wang and Z. Zhang. An integer programming based bound for locally repairable codes. *arXiv preprint arXiv:1409.0952*, 2014.
- [41] A. Wang and Z. Zhang. Repair locality with multiple erasure tolerance. *Information Theory, IEEE Transactions on*, 60(11):6979–6987, Nov 2014.

- [42] A. Wang, Z. Zhang, and M. Liu. Achieving arbitrary locality and availability in binary codes. to appear in *IEEE International Symposium on Information Theory 2015*. Online available: <http://arxiv.org/abs/1501.04264>.
- [43] A. Weiss. Girths of bipartite sextet graphs. *Combinatorica*, 4(2-3):241–245, 1984.
- [44] T. Westerbäck, R. Freij, T. Ernvall, and C. Hollanti. On the combinatorics of locally repairable codes via matroid theory. *CoRR*, abs/1501.00153, 2015.
- [45] A. Zeh and E. Yaakobi. Optimal linear and cyclic locally repairable codes over small fields. *CoRR*, abs/1502.06809, 2015.

SCHOOL OF MATHEMATICAL SCIENCES, CAPITAL NORMAL UNIVERSITY, BEIJING 100048, P.R. CHINA  
*E-mail address:* junz@cnu.edu.cn

SCHOOL OF MATHEMATICAL SCIENCES, ZHEJIANG UNIVERSITY, HANGZHOU 310027, P.R. CHINA  
*E-mail address:* 11235062@zju.edu.cn

SCHOOL OF MATHEMATICAL SCIENCES, CAPITAL NORMAL UNIVERSITY, BEIJING 100048, P.R. CHINA  
*E-mail address:* gnge@zju.edu.cn